UNITED STATES
PATENT AND
TRADEMARK OFFICE

CHIEF INFORMATION OFFICE
UNITED STATES PATENT AND TRADEMARK OFFICE
WASHINGTON, DC 20231
WWW.USPTO.GOV

Office of the Chief Information Officer
Technical Note: IT-212.2-05 TN01
Effective: May 4, 2001
Issuing Office: Office of Data Management

Standard Generalized Markup Language (SGML) and eXtensible Markup Language (XML) Resource Management Guidelines

1. PURPOSE

This technical note establishes procedures for managing Standard Generalized Markup Language (SGML) and eXtensible Markup Language (XML) document resources within the U.S. Patent & Trademark Office (USPTO). The XML Resource Repository stores all versions of DTDs, Schemas, Style Sheets, samples of Document Instances, Public Entities, definitions of USPTO-established XML Namespaces, Templates, and associated documentation. Note: Throughout this document, reference to XML should be construed as including SGML as well.

2. SCOPE AND APPLICABILITY

This technical note applies to all USPTO employees, contractors, and consultants under the direction of the Chief Information Officer. All personnel shall adhere to its contents and the procedures specified herein.

The purpose of this technical note is to avoid unnecessary and costly conflicts in practice among various projects implementing XML and to gain the full benefit of investment in XML technology. The introduction of XML in an organization appears to provide the opportunity for tying automated systems to business rules much more intimately than would be the case without XML. The goal of all interventions in AIS projects described below is to ensure that the interests of all those who have a stake in a system are taken into account when establishing, updating, using, changing, or retiring XML resources.

3. ROLES AND RESPONSIBILITIES

The roles and responsibilities of the XML Registrar, the XML Technology Working Group (TWG) and the AIS System Development Manager are outlined below.

A. XML Registrar (Office of Data Management/Data Administration Division)

As the XML Registrar, the Office of Data Management/Data Administration Division (ODM/DAD) is responsible for controlling access to all approved XML resources. The XML Registrar will conduct internal reviews against logical data models to ensure consistency between logical data elements and the elements used in approved XML resources. The XML Registrar will work with the XML-TWG in comparing new or revised resources against those already approved to ensure maximum re-use of existing XML resources where possible and to assess the impact of any new resources proposed. Upon promotion to the Deployment stage in XML Resource Repository by the SDM with the concurrence of the XML-TWG, resources will be registered as official USPTO named entities and stored in the USPTO Enterprise Information Repository. The Registrar will ensure that XML resources are available to project personnel within the USPTO.

B. XML Technology Working Group

The primary purpose of the XML-TWG is to protect and leverage previous investments in XML resources while supporting the success of new XML projects. The XML Technology Working Group (XML-TWG) will review development of new XML resources, especially XML Schema and DTDs, to ensure conformance with agreed styles and practices; and coordinate changes to resources to maintain consistency across the enterprise and minimize impact on resource users, both internal and external to the organization. Review of XML Schema and DTDs will be a standard LCM step for AIS development projects that include XML resources.

C. AIS System Development Manager (SDM)

The System Development Manager will work with the XML Registrar and the XML-TWG to make use of information already contained in the XML Resource Repository as well as in the USPTO Enterprise Data Model and the USPTO's Standard Data Elements. The System Development Manager will coordinate with the program sponsor regarding XML resources, such as DTDs, schemas, and style sheets.

4. XML Resource Repository Components

The XML Resource Repository identifies the Life Cycle Management (LCM) phases and in addition, users may identify additional stages within the tool. Approved resources are those that have reached the Deployment stage and are ready to enter or have entered the Production stage. Existing projects are encouraged to use the XML Resource Repository for AIS production.

The XML Resource Repository stores all versions of DTDs, Schemas, Style Sheets, samples of Document Instances, Public Entities, definitions of USPTO-established XML Namespaces, Templates, and associated documentation. The approved resources contained in the XML Resource Repository are made available to all project personnel within the PTO via the USPTO Enterprise Information Repository, which is currently available on USPTO workstation desktops and is accessible via the ODM website on the USPTO's Intranet. As part of the configuration management process as defined by the LCM, the AIS development team will provide copies of XML resources, along with other appropriate materials, to the Office of System Product Assurance (OSPA) for loading into the Configuration Management tool.

5. XML RESOURCES MANAGEMENT PROCEDURES

The procedures described here are subject to revision in the light of experiences gained during the months following publication of this Technical Note.

A. XML RESOURCE REPOSITORY PROCEDURES

The following outlines the standard procedures for the XML Registrar, the XML-TWG, and the AIS developers to create and manage XML resources.

(1) Any project developing a system that processes content (patent applications, patent file wrappers, trademark applications, granted patents, published trademark registration certificates, the Manual of Patent Examining Procedure, the Trademark Manual of Examining Procedure, or other similar USPTO-owned documents which record or regulate the official business of the USPTO), and uses XML to do so, will use the XML-TWG and the USPTO's XML Resource Repository as the starting place for XML development. The SDM notifies the XML Registrar or XML-TWG of the new project either at a regular XML-TWG meeting, or by email announcing the project. The XML Registrar or XML-TWG will respond by citing related or similar projects and their associated resources.

(2) SDM requests training in XML Resource Repository from XML Registrar for project personnel.

(3) During the Concept phase of the project life cycle, the SDM establishes categories (the structure used to store resources in XML Resource Repository) based on the development phases of the project.

(4) The SDM determines if the XML resources in XML Resource Repository can be used to fulfill the target system's needs. If not, the SDM contacts the XML-TWG for assistance in modifying existing XML resources or developing new ones. Sharing resources can significantly reduce the cost of developing code to process those resources and improve consistency and reliability of products in the eyes of USPTO customers.

(5) Where existing resources require modification as a consequence of the new project, or new resources have significant consequences for existing projects, the interested parties will evaluate the impacts and negotiate the modifications at one or more XML-TWG meetings.

(6) After completing modifications to existing or developing new XML resources, with the concurrence of the XML-TWG, the SDM advances the resources to the Deployment stage in XML Resource Repository .

(7) The XML resources should be tested during the project's development phase.

(8) When XML resources reach the Production stage in XML Resource Repository , the XML Registrar makes them available for public comments through the USPTO Web site. The selected XML resources will also be available in the Enterprise Information Repository.

(9) Any changes to shared XML resources must be coordinated through the XML-TWG. The goal is to reduce or eliminate unexpected consequences for all users of the resources and to ensure that adequate notice is given so that all systems can be updated to reflect the changes.

(10) XML DTDs, schemas, and other XML resources should be incorporated into a project's Detailed Design Document. For further guidelines, please refer to *the Detailed Design Document Technical Standard Guideline IT-212.4-12*. XML resources shall be added to the Configuration Management tool using the procedures outlined in the *Configuration Management Technical Standard Guideline IT-212.2-06*. It is the SDM's responsibility to incorporate XML resources into an AIS's LCM documentation.

These instructions will be refined as the USPTO gains experience with XML Resource Repository .

B. REQUIREMENTS

1. XML Element Naming Convention

Unabbreviated business terminology should be used as the first choice for naming XML resources. Where applicable, the standard data element name should be used. The *Data Element Standardization - Technical Standard Guideline IT-212.2-13* should be used as a reference. Names should be readily understood by users outside the USPTO and by users inside the USPTO who are not familiar with the AIS under development.

Since SGML is not a case-sensitive markup language and XML is, following industry-wide practice, element names should be all lower case. The total length for an element name should not exceed 64 characters, a convention established by the ODM.

Where a name is composed of multiple words, separate the words with dashes or capitalize the initial letter of interior words. For example, `<title-of-invention>` or `<titleOfInvention>.`

2. XML Resource Naming Convention

a) File Naming Convention

The following guidelines provide the naming conventions for the XML resources created within the USPTO. These conventions should be applied to resources developed in SGML, XML, HTML, or any other SGML-based markup language.

-Begin with the title of the document

-Inclusion of a version number in the file name of production files is deprecated. As a rule, version is determined by directory path, as shown in the examples below.

All file names must be followed by a file-name extension that specifies file type. Some examples of typical XML resources have the following customary extensions:

DTD	Document Type Definition
XSD	XML Schema Definition
SGM	SGML document instance
XML	XML document instance
CSS	Cascading Style Sheet
XSL	eXtensible Style Sheet
XSLT	XSL Transformation
DOC	MS Word document
ENT	SGML/XML Entity Definition
TXT	ASCII text document

There should be no spaces in the file name. Words may be separated by hyphens; please do not use underscores.

-Hypothetical examples of filenames, as they will appear in a publicly accessible FTP site, are:
xml.uspto.gov/mathml/v12a/mathml.dtd
xml.uspto.gov/calstabl/v3/calstabl.dtd
xml.uspto.gov/entities/1986/iso-lat1.ent
xml.uspto.gov/entities/1991/isotech.ent
xml.uspto.gov/rb-grant/test/st32-us-grant-019.dtd
xml.uspto.gov/rb-grant/v2.3/st32-us-grant.dtd
xml.uspto.gov/rb-grant/docs/GrantRedBook.doc
xml.uspto.gov/rb-grant/docs/GrantRedBook-header.txt
xml.uspto.gov/rb-application/v1.9/pap.dtd
xml.uspto.gov/efs/specification/current/u-specif.dtd
xml.uspto.gov/efs/specification/current/u-specif.xsl
xml.uspto.gov/efs/specification/v0.8/u-specif.dtd
xml.uspto.gov/team/efw/test/d-efw.dtd

b) XML Resource Header Conventions

For each XML resource, specific header information should be included that details the contents, purpose, and version of the XML resource file. Each XML resource file should contain information pertaining to who owns the XML resource and from which development organization the XML resource stems. The following guidelines should be used when creating an XML resource header. In an SGML or XML resource, the header will ordinarily be a comment associated with the root element (in the case of DTDs or schema) or the first comment appearing in the file. If the file type does not support internal comments, then the header should be in a separate flat ASCII text file with the same name as the resource with the "-header" appended and the file extension *.txt.

Title	Express the title in words that can be easily understood outside the context of the AIS for which the resource

	was developed. Avoid acronyms and abbreviations.
Version Number (Version numbers in file names are deprecated.)	Add the current version of the file in the format, “vx.xx” where the x’s are replaced by a numerical sequence incremented whenever the file is updated, starting with “0.01” Versions produced before production starts should be numbered “v0.xx.” Subsequent to the start of production, revised versions should be numbered “v1.xx” and higher. Each change to the version number must be recorded in the resource’s Revision History.
Note	A brief explanation of the background and purpose of the resource.
Development Organization Information	Name and address of the organization developing the resource. This will always be U.S. Patent & Trademark Office.
Responsible Party	Give the primary organization name that owns the resources.
Revision History	The first entry in the Revision History should always be: “`Original Version 0.1 - [date]; [responsible party]`”. Add a new entry in the revision history whenever there are changes made, including the date. Within the entry, provide details of each change made. Within each entry, note any increment of the version number.

An example of a DTD header is below.

```
<?xml version="1.0" encoding="UTF-8"?>
<!-- DOCTYPE patent-application-publication [ -->

<!-- DOCUMENT TYPE DEFINITION FOR UNITED STATES PATENT APPLICATION PUBLICATIONS
Reference this DTD as PUBLIC "-//USPTO//DTD PAP V1.3 2000-09-06//EN"
Alias: Application Red Book (ARB)

Development Organization Information: the U.S. Patent & Trademark Office

Note:

The structure is for a patent application electronic document (Pre-grant U. S. publication).
It contains all elements, content, and references to external entities that constitute the
patent application, including bibliographic, abstract, description, sequence listing, claim,
and drawing information.


Responsible Party:

Information Products Division
U.S. Patent and Trademark Office
Crystal Park 3, Suite 441
Washington, DC 20231
-->

<!-- ***** START REVISION HISTORY *****

Pending Issues/Comments:
1)  The CALS Exchange table mode does not support spanning (spanspec) columns.
..Since the table related elements have been removed and replaced with a
..reference to the CALS Exchange XML dtd, this may conflict with ESF
..input. Need to investigate how to implement spanspec within PAP for
..EFS compliance.
```

```
2)  The PAP sample data includes a hard coded reference to the DTD using
..a full path. Need to explore using a URI/URL.

Revised 2000-09-06
1)  Changed version to V1.3 and build date to 2000-09-06.
2)  Changed element usc371-date to optional since it will not be included
..on the published application cover page.
3)  Added optional element military-address to the address element.
..The data capture contractor will extract the military address text
..from the address line text.
4)  Modified the residence element to distinguish between US and
..non-US residences.

Revised 2000-08-22
1) Removed addition-to related application code.
2) Changed version to V1.2 and date to 2000-08-22.

Revised 2000-08-08
1) Changed the xml encoding from "ISO-8859-1" to "UTF-8"

Revised 2000-08-02
1)  Moved the botanic model from the specification (in front of
..the background-of-invention) to the bibliographic (after the
..title-of-invention) section.
2)  Within the patent application DTD, the sequence-cwu model was modified
..to include an object reference for patents with sequence listings that
..are not to be published. The sequence-cwu model changed:
..From: (number,(sequence-list-old-rules | sequence-list-new-rules | table),image*)
..To: (number,(sequence-list-old-rules | sequence-list-new-rules | table | object-
reference),image*)
3) Moved element sequence-cwu to before subdoc-claims.
4) Added a-371-of-international to the patent application DTD (missing)
5) Added reissue-of to the patent application DTD (missing)
5)  Renamed element "substitute-for" to "substitution-for" so it would
..match the EFS counterpart element.
6)  Added footnote to the paragraph content model.
..Moved out of the external references.
8) Changed version to V1.1 and date to 2000-08-02.

. . .

Revised 2000-04-04
..Changed version to 0.2 and date to 2000-04-04.
..Added a new element, number, to the content model of paragraph, biological-deposit
...program-listing-deposit, and sequence-cwu. Used for object sequence number.
...See specification for markup instructions.
..Added a new attribute, inid-code, to almost all the elements in
...subdoc-bibliographic-information. This attribute will contain the INID code as
...assigned by data-capture contractor and which appears on the page-composed
...image next to that element. See specification for details.

Revised 2000-03-27
..Extensive revisions based on PGPub DTD Team meetings over the past two months.
..First official distribution of the DTD on this date.

Created 2000-02-16
********** END REVISION HISTORY ********** -->
```

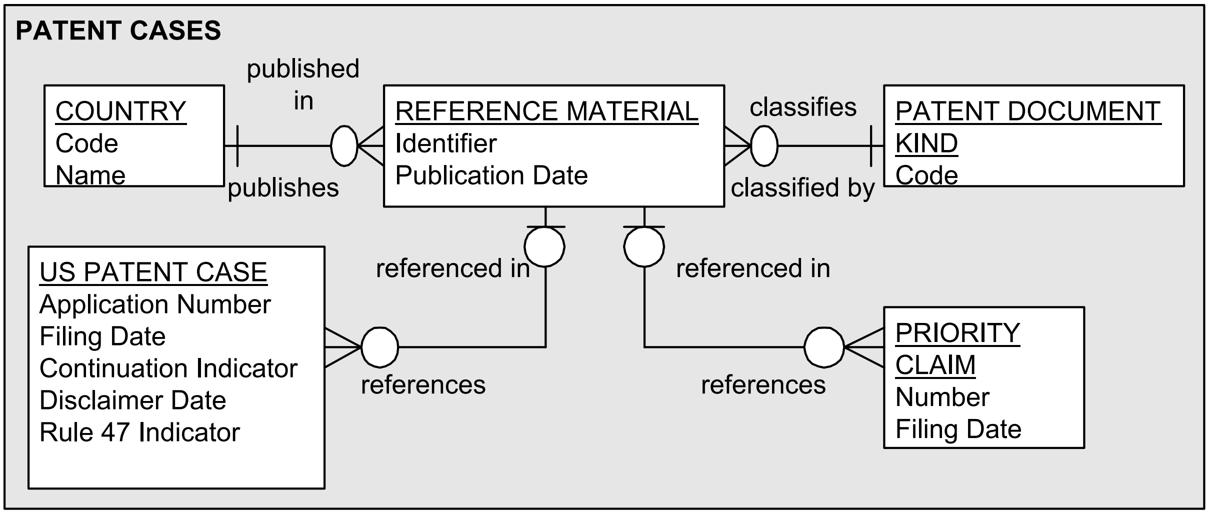
3. Data Modeling Activities

Mapping Matrix

A mapping matrix between elements in a DTD or schema to either the logical data model or to object data in an object-oriented design might be required. Depending on the AIS development approach, these mappings may be produced using traditional mapping documentation (spreadsheets) or via Unified Modeling Language (UML). This matrix will map the transformation from the attributes of the logical data model to the elements of the XML DTD or schema (see example in Figure 2 below). For assistance in preparing the mapping matrix, please contact the Data Administration Division staff.

For AIS's using object-oriented design and UML, please contact the Data Administration Division for assistance on how to map these data elements. Guidelines for developing a logical data model can be found in the *Detailed Design Document (DDD)-Technical Standard Guideline* and *Data Element Standardization-Technical Standard Guideline*.

An example of a logical data model is shown in the below figure.



Legend:

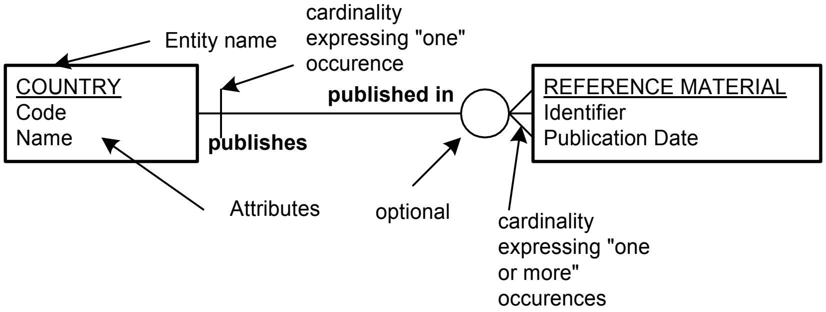


Figure 1—Example of a Logical Data Model

LOGICAL DATA MODEL			DTD/SCHEMA	SCHEMA only	
Attribute Name	***Length***	***Optionality***	***Element Name***	***Length***	***Optionality***
			document-identification		
REFERENCE MATERIAL Identifier	12	Mandatory	doc-number		Mandatory

PATENT DOCUMENT KIND Code	1	Mandatory	kind-code		Mandatory
REFERENCE MATERIAL Publication Date	8	Mandatory	document-date		Mandatory
COUNTRY Code	3	Mandatory	country-code		Mandatory
			domestic-filing-data		
US PATENT CASE Application Number	8	Mandatory	application-number		Mandatory
US PATENT CASE Filing Date	8	Optional	filing-date		Optional
US PATENT CASE Rule 47 Indicator	1	Mandatory	rule-47-flag		Mandatory
US PATENT CASE Continuation Indicator	1	Mandatory	continued-prosecution-application-flag		Mandatory
			foreign-priority-data		
PRIORITY CLAIM Number	10	Mandatory	priority-application-number		Mandatory
PRIORITY CLAIM Filing Date	8	Mandatory	filing-date		Mandatory

Figure 2—Example of A Mapping Matrix

4. Reusable XML Resource Components

The approved XML resources in XML Resource Repository should be reused as much as possible. The SDM is required to implement existing XML resources in XML Resource Repository if they are suitable, or can be modified to be suitable, before developing new ones. Doing so can reduce development time for the resources as well as for software to process the resources.

Appendix A
Guideline for Developing the XML Resources

This appendix provides the recommended guidelines for system developers to create XML resources.

A. EXISTING DOCUMENT ANALYSIS

A thorough analysis of the representative documents and the project requirements is recommended. The business representative(s) should provide the development team with all relevant specifications and reasonable access to an adequate set of documents concerning the type of document for which the XML resource is to be developed. The purpose and outcome of the documentation analysis is to identify structure, format, and content of the given document and to prepare for the development of the XML resources.

The outcome of document analysis should result in the following:

- determine the stakeholders to include users, management, and customers
- identify the scope of the analysis and project (applicability within the organization)
- define a DTD strategy (simple vs. complex, modular vs. in-line, DTD vs. XML Schema)
- identify the root structure of the document class
- identify the logical structures in the document class
- identify the hierarchy of the elements/document tree
- identify specific styling requirements as distinguished from informal styling conventions

B. XML Resource Development

Prior to writing the DTD, it is important to determine the setup of the DTD. It is important to consider:

- If the new XML resource should consist of multiple components; and
- The reusability aspects of each piece of the components if they are to be called or referenced by other resources

It may be useful to decompose the XML resource into components which are stored as external entities. A component is a discrete portion of a DTD or schema that can be reused or called by other XML resources. These may be as small as a single element within that DTD or Schema, or it may be a set of elements combined that create a single logical section of a document type. Such components are often common to more than one document type. For example, the "table" component of a document type contains all the elements necessary to represent the complete structure of a table. The DTD specifying markup for tables (common to any document type that uses tables) is usually in a separate file and called in or accessed through a parameter entity. There are similar DTDs for mathematics markup and for chemistry markup. A DTD which incorporates external entities of this kind is expanded by the parser in a production system to include the declarations contained in the external DTDs. Most DTDs will contain references to external entities which are commonly used within the USPTO. A parameter entity is a pointer to information contained either within the current DTD file or in an external file. Parameter entities may be used to incorporate the content of additional DTDs, components of DTDs, or other files into a DTD. Common components that are maintained as separate, secondary files for the DTD are:

- DTD default structures;
- Special fonts (character entity files);
- Table structures;
- List structures;
- Reference structures;
- Graphic notations; and
- Configuration parameters.

When developing the DTD, case sensitivity must be kept in mind. While SGML is not a case-sensitive markup language, XML is. As such, where possible, element names should be all lower case. Names should be kept to 64 characters or less, with multi-word names underscored (`<first_named_inventor>)`, hyphenated (`<first-named-inventor>`), or use "camel" capitalization (`<firstNamedInventor>`). Hyphenation is preferred as it is the easiest to type.

The XML resources should be validated by at least two parsers. Parsing an XML resource does not determine if it is accurate for the documents for which it is intended but, rather to verify if the source code of the resources is syntactically correct. For the parser products, please refer to the Technical Reference Model.

APPENDIX B

GLOSSARY OF TERMS

The following acronyms and terms were used in this document and are defined here for reference.

TERM	DEFINITION
AIS	Automated Information System
Categories	Categories serve as the primary organizational units within XML Resource Repository. Categories have the same hierarchical structure as directories. Categories usually reflect the development stages as defined in the project.
Component	A DTD or Schema component may be as small as a single element within that DTD or Schema, or it may be a set of elements combined that create a single part of a document. For example, the "table" component of a document contains all the elements and declarations necessary to represent the complete structure of a table. This component ordinarily is in a separate file and is incorporated through a parameter entity.
DAD	Data Administration Division
Document Instance	A document instance is a marked up piece of data that conforms to a DTD or Schema.
Document Type	A document type is a specific genre of document or class of data that can be generalized and to which a general structure can be applied.
DTD	Document Type Definition; used to define the structure and syntax of an XML or SGML document.
HTML	HyperText Markup Language; the language that all web pages were written in prior to the advent of XML. HTML does not delineate between content and format. An element in HTML often represents a specific style or format as opposed to a logical structure.
LCM	Life Cycle Management
LDM	Logical Data Model; the depiction of the relationships and structure of a given class of documents.
Namespace	A namespace indicates where the definition, or meaning of that element can be found. It identifies the source of the element and that element's meaning relative to its source.
ODM	Office of Data Management
OSPA	Office of System Product Assurance
Parameter Entity	A pointer to information contained either within the current DTD file or in an external file. Parameter entities may be

	used to link the content of additional DTDs, components of DTDs, or other files into a DTD.
Public Entities	Public Entities are external to the document instance, schema, or DTD. A public entity is defined using external naming conventions that identify the standard used for the external file, the type of file being referenced (external DTD fragment, character set, etc.), and a unique location or name where that file can be located. With XML, the unique location is often a URL, though in many cases, depending on the longevity required for the identifier, the unique location may be more open. In such cases, a SYSTEM identifier must be mapped to the public identifier.
SDE	Standard Data Element
SDM	System Development Manager (*see Section 3.C*)
Schema	An XML Schema is an XML language for describing and constraining the content of XML document.
SGML	Standard Generalized Markup Language; the “parent” of both XML and HTML; a syntax for structurally and hierarchically representing content independent of its format.
XML Registrar	An ODM staff member responsible for administering and managing all USPTO XML resources.
Style Sheet	A specification of the rendering of a document type or a document instance.
Tag	Tags are text structures that mark the beginning and end of elements within SGML or XML documents. Tags are either paired (start tag and end tag) or empty (only one tag which serves both purposes).
Template	A gauge in creating the XML resource to ensure that data remains accurate, reliable, and consistent.
UML	Unified Modeling Language
URL	Universal Resource Locator
USPTO	United States Patent and Trademark Office
XML	eXtensible Markup Language; the middle “sibling” between SGML and HTML; XML is a profile of SGML, restricting many of the optional aspects of SGML to a single choice, but is more robust and flexible than HTML. XML maintains the delineation between content and format, but removes much of the complexity of SGML.
XML Namespaces	XML Namespaces are used to distinguish specific element names when elements are mixed and matched from different XML applications.
W3C	The World Wide Web Consortium was created in October 1994 to lead the World Wide Web to its full potential by developing common protocols that promote its evolution and ensure its interoperability. W3C has more than 400 Member organizations from around the world and has earned

international recognition for its contributions to the growth of the Web.

[SIGNED]
Holly Higgins
Director, Office of Data Management

May 4, 2001
Date